

KLASIFIKASI DAN PEMBANGKITAN INDEKS UNTUK PENCARIAN
KOLEKSI DOKUMEN ONLINE DENGAN MENGGUNAKAN METODE
NAIVE BAYES CLASSIFIER DAN VECTOR SPACE MODEL

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai gelar Strata Satu
Program Studi Informatika



Disusun oleh:

ARBA SASMOYO

M0511007

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

SURAKARTA

2016

SKRIPSI

KLASIFIKASI DAN PEMBANGKITAN INDEKS UNTUK PENCARIAN
KOLEKSI DOKUMEN ONLINE DENGAN MENGGUNAKAN METODE
NAIVE BAYES CLASSIFIER DAN VECTOR SPACE MODEL

Disusun oleh:

Arba Sasmoyo

M0511007

Diajukan untuk memenuhi sebagian persyaratan memperoleh gelar Strata Satu
Program Studi Informatika

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

SURAKARTA

2016

SKRIPSI
KLASIFIKASI DAN PEMBANGKITAN INDEKS UNTUK PENCARIAN
KOLEKSI DOKUMEN ONLINE DENGAN MENGGUNAKAN METODE
NAIVE BAYES CLASSIFIER DAN VECTOR SPACE MODEL

Disusun oleh:

Arba Sasmoyo

M0511007

Skripsi ini telah disetujui untuk dipertahankan dihadapan dewan penguji,

Pada tanggal: 19 Januari 2016

Pembimbing 1,



Ristu Saptono S.Si., M.T.

NIP. 197902102002121001

Pembimbing 2,



Dr. Wiranto M.Kom., M.Cs.

NIP. 196612301993021001

SKRIPSI
KLASIFIKASI DAN PEMBANGKITAN INDEKS UNTUK PENCARIAN
KOLEKSI DOKUMEN ONLINE DENGAN MENGGUNAKAN METODE
NAIVE BAYES CLASSIFIER DAN VECTOR SPACE MODEL

Disusun oleh:

Arba Sasmoyo

M0511007

Telah dipertahankan dihadapan Dewan Penguji,
pada tanggal: 19 Januari 2016

Susunan Dewan Penguji

1. Ristu Saptono S.Si., M.T.

NIP. 197902102002121001

2. Dr. Wiranto M.Kom., M.Cs.

NIP. 196612301993021001

3. Hasan Dwi Cahyono S.Kom., M.Kom.

NIP. 198205242014041001

4. Sari Widva Sihwi S.Kom., M.T.I.

NIP. 198304122009122003

Disahkan oleh



Kepala Program Studi Informatika,

Dr. Bambang Harjito M.APP.Sc, Ph.D.

NIP. 196211301991031002

HALAMAN PERSEMBAHAN

*Tugas akhir ini ku persembahkan untuk
kedua orang tua dan adik adik ku tercinta,
teman teman informatika angkatan 2011,
keluarga besar UPT TIK UNS.*

MOTTO

“Barang siapa yang menempuh jalan untuk mencari suatu ilmu. Niscaya Allah memudahkannya ke jalan menuju surga”

HR. Turmudzi

“Ya Allah aku memohon kepada-Mu ilmu yang bermanfaat, rizki yang baik, dan amal yang diterima”

HR. Ibnu Majah

“Maka apabila kamu telah selesai (dari sesuatu urusan), kerjakanlah dengan sungguh-sungguh (urusan) yang lain,”

Q.S. Al Insyirah: 8

“The quieter you become the more you can hear”

Baba Ram Dass

“The First Rule of Programming: It's Always Your Fault”

Coding Horror

KATA PENGANTAR

Segala puji penulis panjatkan kehadiran Allah atas limpahan nikmat, hidayah dan inayah-Nya sehingga penulis dapat menyelesaikan tugas akhir yang berjudul “Klasifikasi dan Pembangkitan Indeks Untuk Pencarian Koleksi Dokumen Online Dengan Menggunakan Metode Naive Bayes Classifier dan Vector Space Model”.

Penulis menyadari bahwa tugas akhir ini masih jauh dari kesempurnaan, baik dari segi penulisan maupun materi. Walaupun demikian penulis berharap semoga tugas akhir ini dapat bermanfaat bagi berbagai pihak. Penulis mengucapkan terima kasih kepada semua pihak yang telah meluangkan waktu untuk memberikan bimbingan dan saran sehingga laporan ini dapat berwujud sebagaimana yang diharapkan, terutama kepada:

1. Ayah, Ibu dan segenap keluarga penulis yang telah memberikan kasih sayang, kesabaran, pengorbanan, do'a serta semangat kepada penulis.
2. Bapak Ristu Saptono, S.Si., M.T. dan bapak Dr. Wiranto M.Kom., M.Cs. selaku dosen pembimbing tugas akhir atas kebaikan dan bimbingan selama penyelesaian tugas akhir ini.
3. Para staff dan teman teman maganger UPT TIK UNS yang telah membantu banyak dalam penyelesaian tugas akhir ini.

Surakarta, Januari 2016

Penulis

KLASIFIKASI DAN PEMBANGKITAN INDEKS UNTUK PENCARIAN KOLEKSI DOKUMEN ONLINE DENGAN MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER DAN VECTOR SPACE MODEL

ARBA SASMOYO

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Sebelas Maret

ABSTRAK

Universitas Sebelas Maret memiliki banyak repositori dokumen *online*. Mengelola repositori dengan jumlah banyak tidaklah mudah. Dengan banyaknya jumlah repositori dokumen tersebut justru mempersulit pengguna dalam mencari dokumen. Selain itu, metode pencarian pada beberapa repositori dokumen kurang optimal karena hanya mempertimbangkan judul saja.

Oleh karena itu, pada penelitian ini diajukan sebuah metode untuk mengindeks dan mencari dokumen yang tersebar di beberapa repositori. Terdapat beberapa langkah untuk mengindeks dokumen yang berbeda antara dokumen berbahasa satu dengan bahasa lain. *Naive Bayes Classifier* digunakan untuk mengklasifikasi sebuah dokumen berdasarkan bahasanya. Selanjutnya, pencarian dokumen dilakukan menggunakan algoritma *Vector Space Model*. Proses klasifikasi dan pencarian diuji menggunakan perhitungan *accuracy*, *precision* dan *recall*. Hasilnya, *Naive Bayes Classifier* memiliki *accuracy* 97,62%, *precision* dokumen Indonesia dan Inggris 98,30% dan 95,56%, dan *recall* dokumen Indonesia dan Inggris 95,28% dan 98,17%. Sedangkan *Vector Space Model* memiliki *precision* dan *recall* sebesar 26,59% dan 100%.

Kata kunci: *Naive Bayes Classifier*, Pengindeksan dokumen, Repositori dokumen, *Vector Space Model*.

*CLASSIFICATION AND INDEX GENERATION FOR SEARCHING ON
ONLINE DOCUMENT REPOSITORIES USING NAIVE BAYES CLASSIFIER
AND VECTOR SPACE MODEL*

ARBA SASMOYO

*Department of Informatics, Faculty of Mathematics and Natural Sciences,
Sebelas Maret University*

ABSTRACT

Sebelas Maret University has many online document repositories. Managing many document repositories is not a simple task. As the number of document repository increases, users will have difficulty searching for a document across multiple repositories. Poor searching method on document repository also give users even more bad experiences.

This research propose a method to index and search document which are located accross multiple document repositories. There are some steps to index documents, and some of them are language specific. Naive Bayes Classifier will be used to classify document according to its language. Document searching will use Vector Space Model algorithm. Document classification and searching will be tested using accuracy, precision and recall. The results showed that Naive Bayes Classifier has accuracy 97.62%, precision for Indonesia and English 98,30 and 95.56% and recall for Indonesia and English 95,28% and 98,17%. Meanwhile Vector Space model has precision and recall 26,59% and 100%.

Keywords: Document indexing, Document repository, Naive Bayes Classifier, Vector Space Model.

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGAJUAN.....	ii
HALAMAN PERSETUJUAN.....	Error! Bookmark not defined.
HALAMAN PENGESAHAN.....	Error! Bookmark not defined.
HALAMAN PERSEMBAHAN	v
MOTTO	vi
KATA PENGANTAR	vii
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI.....	x
DAFTAR TABEL.....	xiii
DAFTAR LAMPIRAN.....	xiv
DAFTAR GAMBAR	xv
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Batasan Masalah.....	3
1.4. Tujuan Penelitian	3
1.5. Manfaat Penelitian	3
1.6. Sistematika Penulisan	3
BAB II TINJAUAN PUSTAKA.....	5

2.1. Dasar Teori.....	5
2.1.1. <i>Web Crawler</i>	5
2.1.2. <i>Tokenization</i>	5
2.1.3. <i>Feature Selection</i>	6
2.1.4. <i>Naive Bayes Classifier</i>	7
2.1.5. <i>Stop Words Removal</i>	9
2.1.6. <i>Stemming</i>	9
2.1.6.1. Nazief Adriani <i>Stemmer</i>	10
2.1.6.2. Porter <i>Stemmer</i>	14
2.1.7. <i>Term Frequency</i> dan <i>Inverse Document Frequency</i>	18
2.1.8. <i>Vector Space Model</i>	21
2.1.9. <i>Cosine Similarity</i>	22
2.2. Penelitian Terkait	23
2.3. Rencana Penelitian	25
BAB III METODOLOGI.....	27
3.1. Pengindeksan Dokumen.....	27
3.1.1. Tahap Pengumpulan Data	28
3.1.2. Pengambilan Dokumen dari YaCy.....	29
3.1.3. <i>Tokenization</i>	30
3.1.4. <i>Classification</i>	31
3.1.5. <i>Stop Words Removal</i>	32
3.1.6. <i>Stemming</i>	33

3.1.7. Penyimpanan Data	34
3.2. Tahap Pembuatan Portal Pencarian.....	34
3.2.1. <i>Tokenization Query</i>	35
3.2.2. <i>Stemming Query</i>	35
3.2.3. Penghitungan Nilai <i>Similarity</i>	36
3.2.4. Menampilkan Hasil	36
3.3. Tahap Pengujian.....	36
3.3.1. Pengujian Klasifikasi	37
3.3.2. Pengujian Pencarian	38
BAB IV HASIL DAN PEMBAHASAN	40
4.1. Pengumpulan Data	40
4.2. <i>Training Classification</i>	41
4.3. <i>Testing Classification</i>	43
4.4. <i>Testing</i> Pencarian	44
4.5. Pembahasan.....	44
BAB V KESIMPULAN DAN SARAN.....	48
5.1. Kesimpulan	48
5.2. Saran.....	49
DAFTAR PUSTAKA	50

DAFTAR TABEL

Tabel 2.1. Contoh <i>tokenization</i>	6
Tabel 2.2. Nilai <i>feature set</i>	7
Tabel 2.3. Contoh dokumen.....	8
Tabel 2.4. Hasil penghapusan <i>stop words</i>	9
Tabel 2.5. Tabel aturan Naizef Adriani.....	12
Tabel 2.6. Langkah Algoritma Porter	15
Tabel 2.7. Contoh dokumen.....	19
Tabel 2.8. TF dokumen 1	19
Tabel 2.9. TF dokumen 2	19
Tabel 2.10. TF dokumen 3	19
Tabel 2.11. Hasil normalisasi TF dokumen 1	20
Tabel 2.12. Hasil normalisasi TF dokumen 2	20
Tabel 2.13. Hasil normalisasi TF dokumen 3	20
Tabel 2.14. Nilai IDF	20
Tabel 2.15. Nilai TF-IDF dokumen 1	21
Tabel 2.16. Nilai TF-IDF dokumen 2	21
Tabel 2.17. Nilai TF-IDF dokumen 3	21
Tabel 2.18. Nilai TF-IDF query	21
Tabel 2.19. Nilai <i>cosine similarity</i> antar <i>query</i> dan setiap dokumen.....	23
Tabel 3.1. <i>Confusion Matrix</i> klasifikasi.....	37
Tabel 3.2. <i>Confusion Matrix</i> pencarian.....	38
Tabel 3.3. Daftar <i>query</i> yang digunakan pada pengujian pencarian	39
Tabel 4.1. Daftar kata sama yang dihapus dari <i>feature set</i>	42
Tabel 4.2. Hasil pengujian <i>classification</i>	43
Tabel 4.3. Hasil pengujian pencarian.....	44
Tabel 4.4. Daftar dokumen gagal dikasifikasi	45

DAFTAR LAMPIRAN

Lampiran 1 Daftar Dokumen	52
Lampiran 2 Contoh nilai <i>feature set</i>	59
Lampiran 3 Contoh respon web service YaCy	61
Lampiran 4 Tampilan <i>screenshot</i> aplikasi	62

DAFTAR GAMBAR

Gambar 3.1. Tahapan pengindeksan dokumen	27
Gambar 3.2. Jumlah dokumen yang berhasil dibaca YaCy	28
Gambar 3.3. Url <i>web service</i> YaCy	30
Gambar 3.4. <i>Regular expression</i> untuk <i>tokenization</i>	30
Gambar 3.5. Contoh hasil penggunaan <i>regular expression</i>	31
Gambar 3.6. ERD <i>database</i>	34
Gambar 3.7. Langkah pembuatan portal pencarian	35
Gambar 4.1. Contoh potongan dokumen berbahasa Inggris	40
Gambar 4.2. Potongan kata yang bisa diambil dari dokumen berbahasa Inggris	41
Gambar 4.3. Potongan dokumen berbahasa Indonesia	41
Gambar 4.4 Potongan kata yang bisa diambil dari dokumen berbahasa Indonesia	41
Gambar 4.5. Contoh nilai <i>feature set</i> dokumen.....	42
Gambar 5 Pengambilan dokumen dari <i>web crawler</i>	62
Gambar 6 Daftar dokumen hasil pengambilan dari <i>web crawler</i>	62
Gambar 7 Hasil <i>testing</i> klasifikasi	63
Gambar 8 Hasil <i>testing</i> pencarian	63